

New User Orientation

Webinar presentation, November 9, 2011
NASA Advanced Supercomputing Division

Outline



- **Computing resources available at NAS**
- **Logging in to NAS systems**
- **Transferring files to/from NAS systems**
- **Setting up your module environment**
- **Compiling your code**
- **Running jobs with PBS**
- **Working with PBS**
- **Lustre Best Practices**
- **Storage Best Practices**
- **Summary**

NAS Systems



- **Pleiades: 11,776-node Intel Xeon cluster (as of Nov. 1, 2011)**
processor family: x86_64
 - 5824 Harpertown nodes: 8 cores and 8GB per node
 - 1280 Nehalem nodes: 8 cores and 24GB per node
 - 4672 Westmere nodes: 12 cores and 24GB per node
- **Columbia: 4 large Single-System-Image systems**
processor family: ia64
 - Columbia21: 512 CPUs and 1 TB memory
 - Columbia22: 2048 CPUs and 4 TB memory
 - Columbia[23,24]: 1024 CPUs and 2 TB memory each
- **Lou: 14 PB mass storage system**
processor family: ia64

NAS Systems: Pleiades front-ends



➤ **pfe1, pfe2, ..., pfe12**

- Harpertown nodes: 8 cores, **16 GB/node, 1 GigE network**
- Used for logging in, and interactive work: editing, compiling, submitting jobs, etc.

➤ **bridge1, bridge2**

- Harpertown nodes: 8 cores, **64 GB/node, 10 GigE network**
- Larger memory for pre- or post-processing, viewing graphics (matlab, tecplot, idl, etc.)
- Still running older SLES10SP3 kernel
- Better network for transferring large files (especially to Lou)

➤ **bridge3, bridge4**

- Nehalem-EX nodes: 32 cores, **256 GB/node, 1 GigE network**
- Network to be upgraded to 10 GigE in the near future
- Running the SLES11SP1 kernel, same as pfeX and compute nodes

Requesting an account



Go to <http://www.nas.nasa.gov/hecc> and use the search box

The screenshot shows the NASA HECC website interface. At the top, the NASA logo is on the left, and navigation links for 'NASA Home', 'HEC Program', and 'NAS Division' are on the right. Below this is a horizontal menu with tabs: 'HOME', 'ABOUT HECC', 'RESOURCES', 'SERVICES', 'ACCOUNTS', 'SUPPORT', and a search box containing 'account request form'. The main banner features the text 'HIGH-END COMPUTING CAPABILITY' and 'Computing power to answer NASA's complex science and engineering questions'. Below the banner, the page shows 'HECC Home / HECC Search Results' and 'Search Results'. The search results section indicates 'Results 1 - 10 for account request form. (0.08 seconds)'. The first result is 'Getting an Account' with a date of 'Jun 7, 2011' and a brief description of the account request process. The second result is 'Account Request Form - NAS Division - NASA' with a date of 'Items 7 - 14' and a warning about false or inaccurate information. The third result is 'Getting an Account (Phase II) - HECC Knowledge Base' with a date of 'Apr 28, 2010' and a description of the form and signature requirements. On the right side, there is a 'USER QUICK LINKS' section with links to 'User News', 'System Status', 'Knowledge Base', 'FAQ', 'Get Accounts', and 'New User Orientation'. Below these links, there is a message from NAS Control Room staff providing contact information: '(800) 331-8737, (650) 604-4444, support@nas.nasa.gov'.

NASA Home | HEC Program | NAS Division

HOME ABOUT HECC RESOURCES SERVICES ACCOUNTS SUPPORT account request form

HIGH-END COMPUTING CAPABILITY

Computing power to answer NASA's complex science and engineering questions

HECC Home / HECC Search Results

Search Results

Results 1 - 10 for account request form. (0.08 seconds)

Getting an Account
Jun 7, 2011 ... Submit a NAS Account Request Form. Submit your form by either U.S. mail or fax: Mail: NASA Ames Research Center; NAS Account ...
www.nas.nasa.gov/hecc/accounts/getaccounts.html

Account Request Form - NAS Division - NASA
Items 7 - 14 ... Account Request Form. NASA Advanced Supercomputing Division (5/2011). FALSE OR INACCURATE INFORMATION PROVIDED ON THIS FORM ...
www.nas.nasa.gov/hecc/assets/pdf/Account_Request_Form.doc

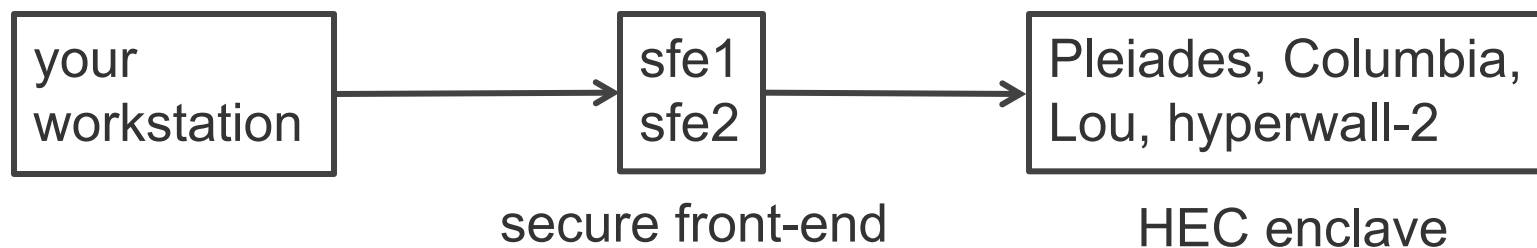
Getting an Account (Phase II) - HECC Knowledge Base
Apr 28, 2010 ... Sign and date the form, and obtain your PI or mission representative signature. 2. Submit the NAS Account Request Form ...
[www.nas.nasa.gov/hecc/.../Getting-an-Account-\(Phase-II\)_51.html](http://www.nas.nasa.gov/hecc/.../Getting-an-Account-(Phase-II)_51.html)

USER QUICK LINKS

- [User News](#)
- [System Status](#)
- [Knowledge Base](#)
- [FAQ](#)
- [Get Accounts](#)
- [New User Orientation](#)

Can't find what you're looking for? NAS Control Room staff are available 24x7x365: (800) 331-8737, (650) 604-4444, support@nas.nasa.gov

Logging in to NAS systems



Two-step connection method: **Easy, but not recommended**

- First, from your wks, login to the secure front-end
`your_wks% ssh sfe1.nas.nasa.gov (or sfe2.nas.nasa.gov)`
or
`your_wks% ssh username@sfe1.nas.nasa.gov`
(if your NAS username is different)
This step requires 8-char pin + passcode from fob and password
(two-factor authentication)
- Second, from sfe1 (or sfe2), login to Pleiades front-end, pfe
`sfe1% ssh pfe`
This step requires password

Logging in to NAS systems



One-step connection method: **preferred method**

`your_wks% ssh pfe`

Enter 8-char pin + passcode from fob

This requires setting up SSH Passthrough

("pass through" because no direct login to sfe1 or sfe2)

The screenshot shows the NASA HECC (High-End Computing Capability) website. At the top, there's a NASA logo and navigation links: HOME, ABOUT HECC, RESOURCES, SERVICES, ACCOUNTS, SUPPORT, and a search bar containing "SSH passthrough". Below the navigation bar is a banner for "HIGH-END COMPUTING CAPABILITY" with the tagline "Computing power to answer NASA's complex science and engineering questions". The main content area shows "Search Results" for "SSH passthrough" with 10 results. The first result is "Setting Up SSH Passthrough - HECC Knowledge Base" dated Oct 22, 2010. The second result is "Setting Up SSH Passthrough" dated May 16, 2011. The third result is "One-Step Connection Using Publickey and Passthrough - HECC ..." dated Jul 21, 2010. On the right side, there's a "USER QUICK LINKS" section with links to User News, System Status, Knowledge Base, FAQ, Get Accounts, and New User Orientation. At the bottom of this section, it says "Can't find what you're looking for? NAS Control Room staff are available 24x7x365: (800) 331-8737, (650) 604-4444, support@nas.nasa.gov".

NASA Home | HEC Program | NAS Division

HOME ABOUT HECC RESOURCES SERVICES ACCOUNTS SUPPORT **SSH passthrough**

HIGH-END COMPUTING CAPABILITY
Computing power to answer NASA's complex science and engineering questions

HECC Home / HECC Search Results

Search Results

Results 1 - 10 for **SSH passthrough**. (0.08 seconds)

Setting Up SSH Passthrough - HECC Knowledge Base
Oct 22, 2010 ... The SSH agent forwarding and an **SSH passthrough** program handle the public key authentication for you, so you will not be prompted for the ...
www.nas.nasa.gov/hecc/support/kb/entry/232

Setting Up SSH Passthrough
May 16, 2011 ... Setting Up **SSH Passthrough**. Category: Security & Logging In. The passthrough feature on the secure front-ends allows you to log into a ...
www.nas.nasa.gov/hecc/support/kb/Setting-Up-SSH-Passthrough_232.pdf

One-Step Connection Using Publickey and Passthrough - HECC ...
Jul 21, 2010 ... This method requires Setting Up Public Key Authentication and Setting Up **SSH Passthrough** first. Once done correctly, use the command ...

USER QUICK LINKS

- [User News](#)
- [System Status](#)
- [Knowledge Base](#)
- [FAQ](#)
- [Get Accounts](#)
- [New User Orientation](#)

Can't find what you're looking for? NAS Control Room staff are available 24x7x365:
(800) 331-8737, (650) 604-4444,
support@nas.nasa.gov

Setting up SSH Passthrough



1. On your workstation:

- Generate public/private key pair
`ssh-keygen -t rsa` (choose a **passphrase**, this command will generate two files: **id_rsa** and **id_rsa.pub**)
- Copy public key to sfe1 (or sfe2)
`scp id_rsa.pub username@sfe1.nas.nasa.gov:~/.ssh2`

2. On sfe1 (or sfe2):

`echo "Key id_rsa.pub" > ~/.ssh2/authorization`

3. On pfe and lou:

- Put contents of id_rsa.pub into ~/.ssh/authorized_keys file
`mv id_rsa.pub ~/.ssh/authorized_keys`

4. On your workstation:

- **Download** the config file from hecc webpage on SSH Passthrough, edit and enter your username, and save the config file under your ~/.ssh directory
- Start ssh-agent
`eval `ssh-agent``
`ssh-add`
(Type your **passphrase** when prompted)

File Systems



- **\$HOME file system is NFS**
 - disk quota: 8GB soft and 10GB hard limit
 - 14 days grace period over soft quota
 - files backed up everyday
- **Scratch directory: /nobackup/userid is a Lustre file system**
 - disk quota: 210GB soft and 420GB hard limit
 - inode quota: 75000 soft and 100000 hard limit
 - 14 days grace period over soft quota
 - files and directories are never backed up

Transferring files to/from NAS systems



Easy if SSH passthrough is already set up

Examples:

- `wks% scp file1 pfe:`
- `wks% scp file1 pfe:file2`
- `wks% scp file1 pfe:dir1`
- `wks% scp -r dir1 pfe:`
- `wks% scp pfe:path_to/file1 .`

Only requires pin + passcode

Use Secure Unattended Proxy to avoid pin+passcode

More cumbersome if SSH passthrough is not set up

- Need to transfer twice, either through `sfe[1,2]` (not recommended, limited disk space) or through `dmzfs[1,2]`
- File transfer cannot be initiated **from** `dmzfs1/dmzfs2` because of their “jailed” environments (limited Unix commands and non-functional `ssh` or `scp` commands). Files can be “pushed” into or “pulled” out of `dmzfs[1,2]`
- Files are automatically deleted from `dmzfs[1,2]` after 24 hours

Transferring files via dmzfs[1,2]



Transferring files from your workstation to Pleiades

- `your_wks% scp file1 dmzfs1.nas.nasa.gov:`
Enter your password (or passphrase if using public key authentication)
- `pfeX% scp dmzfs1:file1 .`
Enter your password
- **8-char pin+passcode from SecurID fob not needed**
- **Cannot initiate file transfer from dmzfs[1,2]**
- **Similarly, you can ssh into dmzfs[1,2], but cannot ssh out**
- **Pulling multiple files from dmzfs[1,2] via wild-card:**
 - `pfeX% scp dmzfs1:* .`
or
 - `pfeX% scp 'dmzfs1:*' .`
- **Using bridge nodes may be faster than pfe for file transfers**

Secure Unattended Proxy (SUP)



SUP Usage Summary:

1. Download and install client (one time)
`your_localhost% wget -O sup http://www.nas.nasa.gov/hecc/support/kb/file/9`
`your_localhost% chmod 700 sup`
`your_localhost% mv sup ~/bin`
2. Authorize directories for writes (one time)
`pfeX% touch ~/.meshrc`
`pfeX% echo /nobackup/jsmith >> ~/.meshrc`
3. Execute command (each time)
`your_localhost% sup scp foobar pfe:/nobackup/jsmith`

Enables direct file transfer from your workstation to Pleiades without going through sfe[1,2] or dmzfs[1,2]

Based on user generating special “SUP keys” with SecurID

- Done automatically via user prompts for passphrase, passcode, password
- **SUP keys are valid for 7 days**

File transfers with **no authentication prompts after generation of SUP keys**

Alternatives to scp



➤ **bbftp**

Pros:

- May provide faster transfer rates than scp
- Can transfer data in parallel using multiple simultaneous streams

Cons:

- Requires downloading and building bbftp client and/or server packages
(Go to <http://www.nas.nasa.gov/hecc> and search for bbftp)
- Command line can be unwieldy:

```
wks% bbftp -V -s -u NAS_username -e 'setnbstream 8; put filename' -E 'bbftpd -s -m 8'  
      bridge1.nas.nasa.gov  (all on one line)
```

➤ **bbscp**

Wrapper script to bbftp provides familiar scp-like syntax:

```
wks% bbscp filename NAS_username@bridge1.nas.nasa.gov:
```

Go to same hecc webpage and search for bbscp to download the script

Use SUP with bbftp/bbscp to avoid multiple authentication prompts

Transfer rates



via scp (**advantage: provides progress status**)

File size	pfe → dmzfs1	bridge2 → dmzfs1	(home) AT&T DSL → dmzfs1
100MB	2s (50.0MB/s)	3s (33.3MB/s)	21m58s (77.7KB/s)
1GB	20s (51.2MB/s)	16s (64.0MB/s)	
10GB	4m56s (34.6MB/s)	3m10s (53.9MB/s)	

via bbscp (**may provide better transfer rates**)

File size	pfe → dmzfs1	bridge2 → dmzfs1
100MB	2s	3s
1GB	12s	13s
10GB	2m59s	4m12s

Setting up module environment



**New account created with no default compilers
(except for GNU compilers)**

`module avail` shows all 172+ modules available
(31 compilers, 19 MPI libraries, 12 HDF5 libraries, etc.)

Recommend adding the following to the end of your .login file:

`module load comp-intel/11.1.072 mpi-sgi/mpt.2.04.10789`

(don't load MKL modules, it's already included in v.11 or later Intel compiler modules)

Default shell is csh (same as tcsh)

Contact control-room if you want a different default shell

Useful module commands:

- `module list` (list currently loaded modules)
- `module purge` (unloads all currently loaded modules)
- `module switch current_module new_module`
- `module show some_module` (shows how your environment variables, PATH, FPATH, LD_LIBRARY_PATH, etc. are changed by loading the module)
- `module help some_module` (info on how some_module was built)

Compiling and Building your code



Intel compilers:

ifort	–	Fortran compiler
icc	–	C compiler
icpc	–	C++ compiler

Compiler options:

aggressive optimization:	-O3 -ip
maintain precision:	-fp-model precise (lowers optimization)
large arrays > 2GB:	-mcmmodel=medium
	-shared-intel (needed at link step)
debugging:	-g -traceback -fpe0 -check

Linking:

MKL math library:	-mkl=sequential
SGL's MPI library:	-lmpi

Example:

```
ifort -c -O3 -ip file1.f90
ifort -c -O3 -ip file2.f90
ifort -o my_exec file1.o file2.o -lmpi
```


Running jobs with PBS



Sample PBS script (run.scr):

```
#PBS -l select=16:ncpus=8:model=har
```

```
#PBS -l walltime=1:00:00
```

```
#PBS -j oe
```

```
cd $PBS_O_WORKDIR
```

```
mpiexec -np 128 ./my_exec > output
```

qsub run.scr

2276977.pbspl1.nas.nasa.gov

qstat -au jsmith (shows all jobs running or queued by user jsmith)

qstat -su jsmith (gives a one line explanation for status of jsmith's jobs)

qstat -nu jsmith (shows nodes used by jsmith's running jobs)

qstat -r (shows all running jobs)

qstat -i (shows all queued jobs sorted by priority)

qdel 2276977 (delete job 2276977)

Running jobs with PBS (continued)



- **‘devel’ queue for faster turnaround (Westmere nodes only)**
 - Can use up to 512 Westmere nodes for 2 hours
 - Each user can run only one job at a time in the devel queue
 - Submit jobs with: `qsub -q devel@pbspl3 run.scr`
`12709.pbspl3.nas.nasa.gov`
 - `qstat -r devel@pbspl3` (shows all running jobs in the devel queue)
 - `qstat -i @pbspl3` (shows all queued jobs served by pbspl3)
- **Interactive PBS jobs (qsub -I)**
 - `qsub -I -lselect=4:ncpus=12:model=wes,walltime=5:00:00`
 - `qsub -I -q devel@pbspl3 -lselect=4:ncpus=12:model=wes,walltime=2:00:00`
 - `qsub -I -v DISPLAY -lselect=4:ncpus=8:model=neh`
`qsub: waiting for job 2277816.pbspl1.nas.nasa.gov to start`
`(Ctrl-c if you don't want to wait)`
 - Default is 1 hour if you don't specify walltime
 - More predictable start time running interactive PBS job in devel queue

Where to run: har or neh or wes?



Time vs. cost considerations

Hypothetical example: running with 128 processes

	# nodes	Walltime (hrs)	SBU rate	Cost (SBU Hrs)
Harpertown	16	20	0.45	144
Nehalem ^(a)	16	11	0.8	140.8
Westmere ^(a)	11	12	1.0	132

(a) Consider running in hyperthreading mode

Queue wait time considerations

- When will my job start to run?
- `node_stats.sh` shows currently queued jobs waiting for which nodes
- `'qstat -W shares -au foo'` gives more detailed info than `node_stats.sh`

Available memory considerations

Harpertown: 8 cores & 8GB/node (rack 32 has 16GB/node)

Nehalem: 8 cores & 24GB/node

Westmere: 12 cores & 24GB/node


HECC Knowledge Base / Memory Usage on Pleiades

http://www.nas.nasa.gov/hecc/support/kb/66

RSS

Google

Apple .Mac eBay Amazon Yahoo! Apple (201) News (3,000) Wireless

NASA Home | HEC Program | NAS Division

HOMEABOUT HECCRESOURCESSERVICESACCOUNTSSUPPORT

HIGH-END COMPUTING CAPABILITY

Computing power to answer NASA's complex science and engineering questions

Knowledge BaseNewsDownloadsAsk a Question

Search: [Advanced search](#)

Knowledge Base

- New User Orientation
- FAQ
- Updates on Issues
- Policies
- Troubleshooting
- Tips & Tricks
- The HEC Environment
- Computing at NAS
 - Computing Hardware
 - Porting & Developing ...
 - Software Environment
 - Running Jobs with PBS
 - Best Practices
 - Effective Use of Re...
 - Memory Usage on ...**
 - Lustre on Pleiades
- Data Storage & Transfer
- Best Practices

HECC Home / Support Home / KB Home / Computing at NAS / Best Practices / Memory Usage on Pleiades

Memory Usage on Pleiades

Options

- Memory Usage Overview
- Checking memory usage of a batch job using qps
- Checking memory usage of a batch job using qtop.pl
- Checking memory usage of a batch job using qsh.pl and "cat /proc/meminfo"
- Checking memory usage of a batch job using gm.x
- Checking if a Job was Killed by the OOM Killer
- How to get more memory for your job

Add /u/scicon/tools/bin to your \$path to use these tools

Using qtop.pl to monitor memory usage



pfe4% qtop.pl 2277539

r193i0n3

top - 01:18:54 up 19 days, 11:32, 0 users, load average: 11.99, 11.97, 11.91

Tasks: 462 total, 13 running, 449 sleeping, 0 stopped, 0 zombie

Cpu(s): 26.4%us, 4.5%sy, 0.0%ni, 69.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st

Mem: 24056M total, 13286M used, 10770M free, 0M buffers

Swap: 0M total, 0M used, 0M free, 8510M cached

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
31311	jsmith	20	0	23.9g	277m	37m	R	101	1.2	171:14.60	overflowmpi
31313	jsmith	20	0	23.9g	274m	36m	R	101	1.1	171:14.55	overflowmpi
31314	jsmith	20	0	23.9g	275m	36m	R	101	1.1	171:14.54	overflowmpi
31315	jsmith	20	0	23.9g	273m	35m	R	101	1.1	171:13.96	overflowmpi
31319	jsmith	20	0	23.9g	272m	36m	R	101	1.1	171:14.32	overflowmpi
31320	jsmith	20	0	23.9g	273m	36m	R	101	1.1	171:14.31	overflowmpi
31309	jsmith	20	0	24.5g	821m	39m	R	99	3.4	171:08.90	overflowmpi
31310	jsmith	20	0	23.9g	275m	37m	R	99	1.1	171:14.61	overflowmpi
31312	jsmith	20	0	23.9g	276m	37m	R	99	1.2	171:14.56	overflowmpi

Lustre Best Practices



Pleiades scratch directory, /nobackup/jsmith, is a Lustre filesystem

/nobackup/jsmith is a symlink to the actual directory: `pfeX% ls -l /nobackup/jsmith`

```
lrwxrwxrwx 1 root root 18 Jul 19 16:53 /nobackup/jsmith -> /nobackupp1/jsmith/
```

Checking quotas on Lustre

```
pfeX% lfs quota -u jsmith /nobackupp1
```

Disk quotas for user jsmith (uid xxxx):

Filesystem	kbytes	quota	limit	grace	files	quota	limit	grace
/nobackupp1	97757456	210000000	420000000	-	42573	75000	100000	

File striping (if and only if file is greater than 1GB)

`lfs setstripe -c 16 -s 4m bigfile` (Sets stripe count of 4 and stripe size of 4MB for bigfile;
must be done before bigfile is created)

`lfs gestripe bigfile` (get information on file striping for bigfile)

`lfs setstripe -c 16 -s 4m bigdir` (sets striping for directory bigdir; all new files created under bigdir will retain the file striping characteristics of bigdir)

Default file striping is -c 1 -s 4m

Lustre Best Practices (cont.)



Avoid repetitive or continuous file *stats* by adding *sleep*

For example, if checking for the presence of file “GO,” instead of:

```
while (! -e GO)
end
```

use

```
while(! -e GO)
sleep 2
end
```

For more on Lustre Best Practices, go to <http://www.nas.nasa.gov/hecc> and search for Lustre. Start with “Lustre Basics.”

Storage Best Practices



- **To find out whether your mass storage system is lou1 or lou2, log into either one and run 'mylou'**
- **Currently, no space quota, but there is a 250,000/300,000 soft/hard limit on number of files (inode quota), with a grace period of 14 days**
- **Files greater than 1MB are migrated to tape**
- **Use 'dmls -l' to see if the files are online on disk (REG), offline on tape (OFL), both online and offline (DUL), unmigrating from tape to disk (UNM), or migrating from disk to tape (MIG)**
- **Use 'dmget filename' to retrieve filename from tape before transferring file**
- **Use 'dmput -rw filename' to release the file from disk and migrate it to tape (if it is not already on tape)**

Storage best practices (cont.)



What's wrong with transferring a bunch of files with:

- (1) `lou% scp *.dat pfe:/nobackupp1/jsmith` or
- (2) `lou% scp -r projectdir remote_host:/nobackup/jsmith ?`

(Answer: It involves repeated loading and unloading of tapes, which is bad for the tape drive and causes lou to slow down for everyone.)

It's better to replace (1) with:

```
lou% dmget *.dat &  
lou% scp *.dat pfe:/nobackupp1/jsmith
```

And (2) with:

```
lou% cd projectdir  
lou% dmfind . -state OFL -print | dmget &  
lou% scp -r ../projectdir remote_host:/nobackup/jsmith  
lou% dmfind . -state DUL -print | dmput -rw
```

For more on Storage Best Practices, go to <http://www.nas.nasa.gov/hecc> and search on 'storage.' See, in particular, the section on "Dealing with Slow File Retrieval"

Summary



- Download account request form**
- Complete NASA Basic IT Security Training**
- Logging in to NAS Systems**
- Set up SSH Passthrough**
- Transferring files via one-step method**
- Transferring files via DMZ**
- Secure Unattended Proxy**
- Using bbftp and bbscp to transfer files**
- Porting and Developing Applications**
- Running jobs with PBS**
- Commonly Used PBS Commands**
- Lustre File Striping**
- Lou best practices**

Miscellaneous



Useful items to add to your .cshrc file

```
set prompt="`hostname -s`:`pwd`> "
```

```
alias cd 'cd \!* ; set prompt="`hostname -s`:`pwd`> "'
```

```
set path=($path /u/scicon/tools/bin .)
```

```
alias qstat 'qstat -W shares'
```

```
alias qstat_m '/u/scicon/tools/bin/qstat -W shares'
```

```
alias ls '/bin/ls -CF'
```

Services Provided at NAS



- **Control-room 24x7**
support@nas.nasa.gov
(650) 604-4444
(800) 331-USER
- **Scientific Consultants M-F**
- **WAN Network support**
- **Visualization**



Slides prepared by Johnny Chang